Fast Evaluation of Appointment Schedules for Outpatients in Health Care

S. De Vuyst, H. Bruneel, and D. Fiems

SMACS* Research Group, Department of Telecommunications and Information Processing, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
{sdv,hb,df}@telin.ugent.be

Abstract. We consider the problem of evaluating an appointment schedule for outpatients in a hospital. Given a fixed-length session during which a physician sees K patients, each patient has to be given an appointment time during this session in advance. When a patient arrives on its appointment, the consultations of the previous patients are either already finished or are still going on, which respectively means that the physician has been standing idle or that the patient has to wait, both of which are undesirable. Optimising a schedule according to performance criteria such as patient waiting times, physician idle times, session overtime, etc. usually requires a heuristic search method involving a huge number of repeated schedule evaluations. Hence, the aim of our evaluation approach is to obtain accurate predictions as *fast* as possible, i.e. at a very low computational cost. This is achieved by (1) using Lindley's recursion to allow for explicit expressions and (2) choosing a discrete-time (slotted) setting to make those expression easy to compute. We assume general, possibly distinct, distributions for the patient's consultation times, which allows us to account for multiple treatment types, as well as patient no-shows. The moments of waiting and idle times are obtained. For each slot, we also calculate the moments of waiting and idle time of an additional patient, should it be appointed to that slot. As we demonstrate, a graphical representation of these quantities can be used to assist a sequential scheduling strategy, as often used in practice.

1 Introduction

1.1 Situation

Because of its social and economic interest, the question of how to schedule a hospital's outpatients into the consultation session of a physician has received a lot of attention over the last sixty years. Many studies are motivated from a specific practical situation and aim at improving the organisational procedures in a particular (part of a) hospital [2,11,19,23]. Clearly, practical settings considerably differ in terms of medical practice, organisation, regulations, administrative demands or limitations, preferences of patients or medical staff, management issues, etc. However, very often the underlying problem is largely the same and

* SMACS: Stochastic Modeling and Analysis of Communication Systems.

K. Al-Begain et al. (Eds.): ASMTA 2011, LNCS 6751, pp. 113–131, 2011.

can be formulated as follows. Consider the practice of a physician who consults patients during a time interval of a certain length called a session, for example a 4-hour session from 8am to 12am every week day. The physician is assisted by a nurse or secretary at the administration desk who is responsible for taking the calls of patients who wish to see the physician during the session of a particular day. The administrator must decide whether a calling patient can be admitted to that session and if so, at what time during the session the patient should arrive, i.e. what is his appointment time. All appointments are fixed before the session starts. The physician arrives at some point during the session, which is not necessarily the beginning. Given the session lengths and the number of patients, a 'schedule' consists of both the patient's appointment times and the physician's arrival time.

How a session evolves depends on its schedule. Since patients are consulted one by one in their appointed order, the patients in the waiting room behave as a FIFO (First-In First-Out) queueing system with the physician as service facility. The time required to serve a single patient is the consultation time, comprising all actions by the physician devoted only to that patient such as examination, looking up test results, giving advice, writing prescriptions, updating files, discussions, etc. It is clear that prior to the session, consultation times are known stochastically only and can be assumed independent. The arrival process on the other hand is not stochastic but consists of scheduled patient arrivals at deterministic time points. Hence, evaluating a session amounts to the study of a queueing system conditioned on a certain sample path for the arrivals. In fact, queueing systems with scheduled arrivals are known as appointment systems [12]. A patient arriving to the session at its appointed time can encounter two possible situations: either the physician has finished the consultations of previous patients or he has not. In the former case the physician has been without work, wasting time, since the departure of the last patient, whereas in the latter case it is the new patient who has to wait. As such, for each appointment there is either an idle time for the physician or a waiting time for the patient. As long as there is uncertainty on the consultation times when making the schedule it is impossible to avoid both idle and waiting times, although they can be controlled to a large extent by the schedule. Note that there is an 'obvious' trade-off. Scheduling appointments far apart results in low waiting times but long idle times and vice versa if the appointments are close together. The same consideration can be made at the end of the session: if the physician has finished all consultations before the end of the session, there is an undertime, whereas otherwise he has to work overtime. Again, session undertime and overtime are antagonistic and to some extent controllable by the schedule.

1.2 Modelling Issues

Depending on the specific situation, there are several so-called *environmental* factors that can make modelling the appointment systems considerably more complex, see [5] for an elaborate discussion. Patients may show up during the session that have no appointment ('walk-ins') but have to be seen by the physician

anyway, either immediately (emergencies), in between regular patients or at the end of the session. Conversely, some patients that have an appointment do not show up for their consultation ('no-shows') or cancel the appointment too late. The no-show probability in some cases is up to 30%, depending on the type of health care offered and the patient population [10,15,21]. Clearly, walk-ins and no-shows contribute significantly to respectively the waiting and idle times of the schedule and to its overall uncertainty. Additionally, patients are not always punctual, for example arriving to the session later or sooner than they are supposed to. According to [1] the difference between appointed and actual arrival time is best modelled by an asymmetric Johnson distribution. Depending on the particularities of the used waiting-room policy, unpunctuality can result in overtaking of patients so that the original order of consultations is no longer maintained. With regard to scheduling, a complicating factor is also the fact that many patients have particular constraints concerning their appointment time. It is reported that as much as 25% of the calling patients [20] ask to be given an appointment in a certain subset of the session.

As to which distribution is suitable for modelling patient consultation times, several propositions have been made. Originally [4,3] Gamma distributions were used, also preferred in e.g. [7]. Other proposed distributions are Cox-type [22], lognormal [6], Weibull [2], uniform and/or exponential [12,13,14,17] and even deterministic consultations [10]. However, patients may also be considered heterogeneous, i.e. have different consultation time distributions. Unlike walk-ins and no-shows, heterogeneity can reduce schedule uncertainty if properly taken into account. For each calling patient, the administration can estimate the required consultation time distribution based on the person's characteristics (age, medical record) and required type of medical treatment (medical scans, surgical procedures, inoculations, revalidation therapy, in-takes, discussion of test results, etc.).

1.3 Schedule Optimalisation

Constructing a schedule is targeted at striking an equitable trade-off between several performance criteria of the schedule such as waiting times of the subsequent patients, physician idle times, session overtime and undertime. Also, more subtle performance issues have been considered to be of importance, such as fairness (uniformity of patient waiting times), the number of patients seen in a session, the degree in which patient constraints can be met, etc. In general, it is not possible to construct the optimal schedule from the desired objective directly. Instead, a search method is required such as sequential quadratic programming [12], modified conjugate direction methods [22], stochastic linear programming [9] or local search methods [14]. These methods all basically work in the same way: take some initial schedule, evaluate it and based on its performance and the objective function try to improve it. Then do the same with the new schedule and so on until it is decided that no more significant improvements can be made. Unfortunately, only in some specific cases can convexity be proven, see e.g. [14]. In any case, since optimalisation requires a huge number of evaluations, it is very important to use an evaluation method that is both accurate and fast.

Concerning optimalisation however, a distinction needs to made between two possible ways of deciding the schedule. In many practical situations, *sequential* scheduling is employed where the schedule is built gradually over time, fixing the appointment for each patient immediately when they call in, until the session is full. With *advance* scheduling on the other hand, the appointment times are optimised for all patients at the same time, which is a much more complex task but can lead to better schedules.

Most studies impose certain limitations on the way appointments can be made and on how a session is organised, either to assure tractability of the evaluation method, reduce the search space or to make practical implementation easier. For example, often a session is be divided in *blocks* (possibly of different length) such that patients can only arrive at the start of a block, see e.g. [7,18,20]. Several scheduling 'rules' have been proposed to determine suitable appointment times for the patients, many of which are summarised in [5]. These rules differ e.g. in the prescribed number of patients in subsequent blocks, initial block size, the length of the intervals between the blocks (either fixed or variable), and so on. In [3] Bailey's recommendation was to have the intervals be equal to the expected consultation time and let the physician start with the second patient. This is now known as 'Bailey's rule' and was aimed at an equitable minimization of both patient waiting and physician idle times. More advanced rules exploit knowledge about patient heterogeneity, i.e. the fact that they have different known consultation time distributions. In [6] for example, a distinction is made between long and short consultations corresponding to 'new' and 'return' patients respectively. In [13] it is shown that it is beneficial to increase the intervals proportional to the standard deviation of each consultation time. Additionally, it is generally better to schedule consultations with low variance early in the session, see [20].

1.4 Discrete-Time Model and Assumptions

In this paper, we propose an analytic schedule evaluation method based on the recursive Lindley relation [16] in queueing theory. Our primary aim is to obtain expressions for the moments of the schedule's performance criteria having very low computational complexity. Key to our approach is the discrete-time setting. That is, not only the session but also all time-related quantities in the model, such as waiting and idle times, are discretised into fixed-length intervals (*slots*) of length Δ. A suitable choice of Δ follows from a trade-off: whereas using small slots ensures a maximal accuracy of the performance predictions, choosing large slots results in a lower computational effort. In the envisaged medical context of appointments for outpatients, a practical time granularity is probably in the order of $\Delta = 1$ minute, as it does not make sense to give people an appointment time with greater accuracy than this. More importantly however, any quantitative description of the consultation time of a patient or of a certain treatment type will rarely require a time granularity smaller than one minute. That is, in as far as the distribution of the anticipated consultation time S is not already made available as discrete data (a histogram), its distribution function can be quantised as

$$s(n) = \text{Prob}[S < (n+\frac{1}{2})\Delta] - \text{Prob}[S < (n-\frac{1}{2})\Delta]\,,\ n \geqslant 0\,, \tag{1}$$

into the probability mass function (pmf) of a *discrete* random variable. Assuming that time is discrete simplifies the analysis considerably, since the integrals over finite intervals that follow from a continuous-time transient queueing analysis (see e.g. [7]) are replaced by finite sums. On the other hand, the discrete-time setting hardly compromises accuracy if the slot length Δ is chosen sufficiently small.

In our analysis, no assumptions are made on the consultation times of the patients other than that their pmf (1) is known and that they are independent. The fact that each patient can have a different consultation pmf allows us to evaluate schedules containing heterogeneous patients, which is important when making tight schedules with low cost. It is clear that the better the consultation time of a patient can be estimated beforehand, i.e. the smaller $\text{Var}[S]$, the better the performance of the optimal schedule will be. For example, almost nothing can be assumed about a new patient seeing the physician for the first time, so a high-variance distribution of S must be assumed. For a patient only needing a prescription for a diagnosed chronic affliction however, the consultation time is almost deterministic. In appointment scheduling, there is much to be gained from a well-considered estimation of the anticipated consultation time of each particular patient. Therefore, the administrator is challenged to use as much advance knowledge about the patient as possible in order to maximally reduce the uncertainty on S. This can for example be done based on the patient's medical history, on time measurements of previous consultations or simply by asking the patient some questions when he calls for an appointment.

We assume that all patients and the physician are punctual, arriving precisely when scheduled. Although patient lateness is excluded in our model, no-shows and even emergencies or other physician unavailabilities can be incorporated to some extent. Specifically, if a patient with consultation time pmf $s(n)$ is likely not to show up for his appointment with probability p, his 'effective' consultation time has pmf

$$s_{\text{no-show}}(n) = \begin{cases} p + (1-p)s(0)\,, & n = 0\,, \\ (1-p)s(n)\,, & n > 0\,. \end{cases} \tag{2}$$

Likewise, if there is a probability q that a consultation with pmf $s(n)$ will be interrupted by an emergency taking a length of time with pmf $u(n)$, then the altered pmf is

$$s_{\text{emergency}}(n) = (1-q)s(n) + q(s * u)(n)\,, \qquad n \geqslant 0\,, \tag{3}$$

where $*$ denotes the discrete convolution. Finally, the physician doesn't necessarily start seeing patients at the start of the session. To anticipate no-shows or lateness of the first few patients, the physician's arrival may be scheduled later in the session.

2 Evaluation of an Appointment Schedule

2.1 Model Description

Consider a consultation *session* of a physician spanning a time period $[0, t_{\max}]$ in which K patients are given an appointment. Let τ_k denote the appointment time of the kth patient ($0 \leqslant \tau_1 \leqslant \tau_2 \leqslant \ldots \leqslant \tau_K \leqslant t_{\max}$) and let θ ($0 \leqslant \theta \leqslant t_{\max}$) be the arrival time of the physician. All patients are assumed to be punctual and their consultation times constitute a sequence of independent random variables, denoted by S_k, $1 \leqslant k \leqslant K$. As already motivated, we assume time to be a discrete dimension where events can only happen at slot boundaries. Therefore, all time related measures are expressed as integer multiples of the slot length Δ. That is, the session length $t_{\max}$, the physician arrival time θ and the appointment times τ_k are given as discrete values and the consultation times S_k have a discrete distribution with pmf $s_k(n) = \text{Prob}[S_k = n]$ which may be obtained from (1) or otherwise available. We denote by μ_k and σ_k^2 respectively the mean and variance of the kth patient's consultation time.

A specific appointment *schedule* thus consists of the session length $t_{\max}$, the physician arrival time θ, the number of patients K, their appointment times τ_k and the consultation time distributions $s_k(n)$. Such a schedule can be evaluated in terms various criteria, among which the patient waiting time and the physician idle time are probably the most important. The waiting time W_k of the kth patient in the schedule is the time between its appointed arrival time and the effective start of its consultation. By the idle time I_k we mean the period *before* the arrival of patient k in which the physician has nothing to do because the consultation of patient $k-1$ is already finished. Usually, for decision-making or optimalisation it is sufficient to predict the mean and the variance of these distributions, which can be calculated very efficiently as we demonstrate.

We denote the interarrival time between consecutive patients by $a_k = \tau_{k+1} - \tau_k$ for $k = 1, 2, \ldots, K$, where it is agreed that $\tau_{K+1} = t_{\max}$ indicates the end of the session. Hence, a_K is the time between the last appointment and the end of the session. We can also interpret $\tau_{K+1} = t_{\max}$ as the arrival time of an additional *virtual* patient at the end of the session. This is useful, since the waiting time of this virtual patient equals the session overtime X, i.e. the excess time beyond the scheduled end of the session that the physician requires to see all K patients. Clearly, the overtime $X = W_{K+1}$ is an important criterium for the performance of the schedule as well.

2.2 Analysis

If we define the auxiliary variable

$$Q_k = W_k + S_k - a_k \,, \tag{4}$$

for $k = 1, \ldots, K$, then the well-known Lindley equation in queueing theory [16] relates the waiting and idle times of consecutive patients as

$$W_{k+1} = (Q_k)^+ \,, \quad \text{and} \quad I_{k+1} = (-Q_k)^+ \,, \tag{5}$$

where $(\cdot)^+$ is a shorthand notation for $\max(\cdot, 0)$. Note that W_{k+1} and I_{k+1} cannot both be positive at the same time, although $W_{k+1} = I_{k+1} = Q_k = 0$ may occur when the consultation of patient k finishes exactly in the slot before the arrival of patient $k+1$. For further calculations, we distinguish between the case $Q_k = -I_{k+1} \leqslant 0$ where patient $k+1$ can be seen immediately and the case $Q_k = W_{k+1} > 0$ where this patient has to wait. In particular, the probability mass function $w_{k+1}(n) = \text{Prob}[W_{k+1} = n]$, $n \geqslant 0$ of the $(k+1)$th waiting time can be related to that of the previous patient using (5). We find

$$\begin{aligned} w_{k+1}(0) &= \sum_{m=0}^{a_k} \sum_{n=0}^{a_k-m} s_k(m) w_k(n)\,, \\ w_{k+1}(n) &= \sum_{m=0}^{n+a_k} s_k(m) w_k(n-m+a_k)\,, \quad n>0\,, \end{aligned} \tag{6}$$

where we exploited the fact that the waiting time of a patient and his consultation time are independent. These probabilities are easy to calculate due to the discrete-time modelling. The first patient is scheduled either before or after the physician's arrival, and has deterministic waiting and idle times respectively given by

$$W_1 = (\tau_1 - \theta)^+\,, \quad \text{and} \quad I_1 = (\theta - \tau_1)^+\,. \tag{7}$$

Hence, the pmf $w_1(n)$ is immediately given and the relations (6) allow us to calculate $w_k(n)$ recursively for all n and k, as far as necessary.

In principle, if the consultation times are bounded, it is possible to calculate the complete probability mass function of the waiting times from which moments can be determined. Such an approach however, is computationally demanding and not applicable if consultation times have unbounded support. Nevertheless, calculation of the probabilities (6) can be partially *avoided* as long as only the moments of waiting and idle times are required. For example, again by (4)–(5), the mean waiting times of subsequently scheduled patients are related as

$$\begin{aligned} \mathrm{E}[W_{k+1}] &= \mathrm{E}[Q_k^+] = \mathrm{E}[Q_k\,\{Q_k>0\}] = \mathrm{E}[Q_k] - \mathrm{E}[Q_k\,\{Q_k\leqslant 0\}] \\ &= \mathrm{E}[W_k] + \mu_k - a_k + \bar{\ell}_k\,, \end{aligned} \tag{8}$$

for $k=1,\ldots,K$ and where $\bar{\ell}_k$ is the finite sum,

$$\bar{\ell}_k = \sum_{m=0}^{a_k} \sum_{n=0}^{a_k-m} (a_k-n-m) s_k(m) w_k(n)\,. \tag{9}$$

In a similar way, we obtain for the waiting time variances

$$\mathrm{Var}[W_{k+1}] = \mathrm{Var}[W_k] + \sigma_k^2 + \bar{\ell}_k^2 - 2\bar{\ell}_k \mathrm{E}[W_{k+1}] - \bar{\bar{\ell}}_k\,, \tag{10}$$

with

$$\bar{\bar{\ell}}_k = \sum_{m=0}^{a_k} \sum_{n=0}^{a_k-m} (a_k-n-m)^2 s_k(m) w_k(n)\,. \tag{11}$$

Again, because of (7) we have that $\mathrm{E}[W_1] = (\tau_1 - \theta)^+$ and $\mathrm{Var}[W_1] = 0$ respectively, such that by (6)–(11) the mean and variance of the waiting times of the patients can be determined recursively for $k = 2, \ldots, K$. It is now also clear that only a finite number of waiting time probabilities need to be calculated by means of (6), even though the consultation times S_k may be stochastically unbounded. Specifically, in accordance with (8)–(9), the calculation of $\mathrm{E}[X] = \mathrm{E}[W_{K+1}]$ requires probabilities $w_K(0) \rightarrow w_K(a_K)$, which in turn requires $w_{K-1}(0) \rightarrow w_{K-1}(a_K + a_{K-1})$ and so on, until finally for the first patient we need $w_1(0) \rightarrow w_1(t_{\max} - \tau_1)$. Note that because W_1 is deterministic, all probabilities in the latter range are zero, except for one. For the variances of the patient waiting times, the same finite set of probabilities

$$\mathcal{W} = \{w_k(n) : 1 \leqslant k \leqslant K, 0 \leqslant n \leqslant t_{\max} - \tau_k\}\,. \tag{12}$$

is used. This set $\mathcal{W}$ is computationally the most demanding part of the schedule's evaluation, in terms of the required number of floating-point multiplications, given by

$$\mathrm{FPM}(\mathcal{W}) = \frac{1}{2}\sum_{k=1}^{K}(t_{\max} - \tau_k + 2)(t_{\max} - \tau_k + 1)\,, \tag{13}$$

in the worst case where the consultation times have infinite support. For a session of length $t_{\max}$ with K patients scheduled at equal distances, $\mathrm{FPM}(\mathcal{W})$ is $Kt_{\max}^2/6 + \mathrm{O}(t_{\max}^2)$.

The moments of the physician idle times I_k that occur before each patient's appointment are related to the moments of the waiting times by means of

$$W_{k+1} - I_{k+1} = Q_k\,, \quad \text{and} \quad W_{k+1}^2 + I_{k+1}^2 = Q_k^2\,, \tag{14}$$

a direct consequence of (5). Hence, for $k = 1, \ldots, K$ one finds

$$\mathrm{E}[I_{k+1}] = \mathrm{E}[W_{k+1}] - \mathrm{E}[W_k] - \mu_k + a_k = \bar{\ell}_k\,, \tag{15}$$

and, since $\mathrm{Var}[Q_k] = \mathrm{Var}[W_k] + \sigma_k^2$ due to (4),

$$\begin{aligned}\mathrm{Var}[I_{k+1}] &= \mathrm{E}[Q_k^2 - W_{k+1}^2] - (\mathrm{E}[W_{k+1}] - \mathrm{E}[Q_k])^2 \\ &= \mathrm{Var}[W_k] - \mathrm{Var}[W_{k+1}] + \sigma_k^2 - 2\mathrm{E}[W_{k+1}]\mathrm{E}[I_{k+1}] \\ &= \bar{\bar{\ell}}_k - \bar{\ell}_k^2\,,\end{aligned} \tag{16}$$

which are all known quantities at this point. Recall that $X = W_{K+1}$ is the session overtime of which mean and variance follows from the algorithm explained above. In the same way, the idle time I_{K+1} associated with the virtual patient at the end of the session can be interpreted as the session undertime, i.e. the time by which the physician finishes the session early after seeing all K patients.

2.3 Examples

In the following examples, we evaluate some particular schedules with regard to the incurred mean waiting, idle and overtime. We assume a slot length of $\Delta = 1$

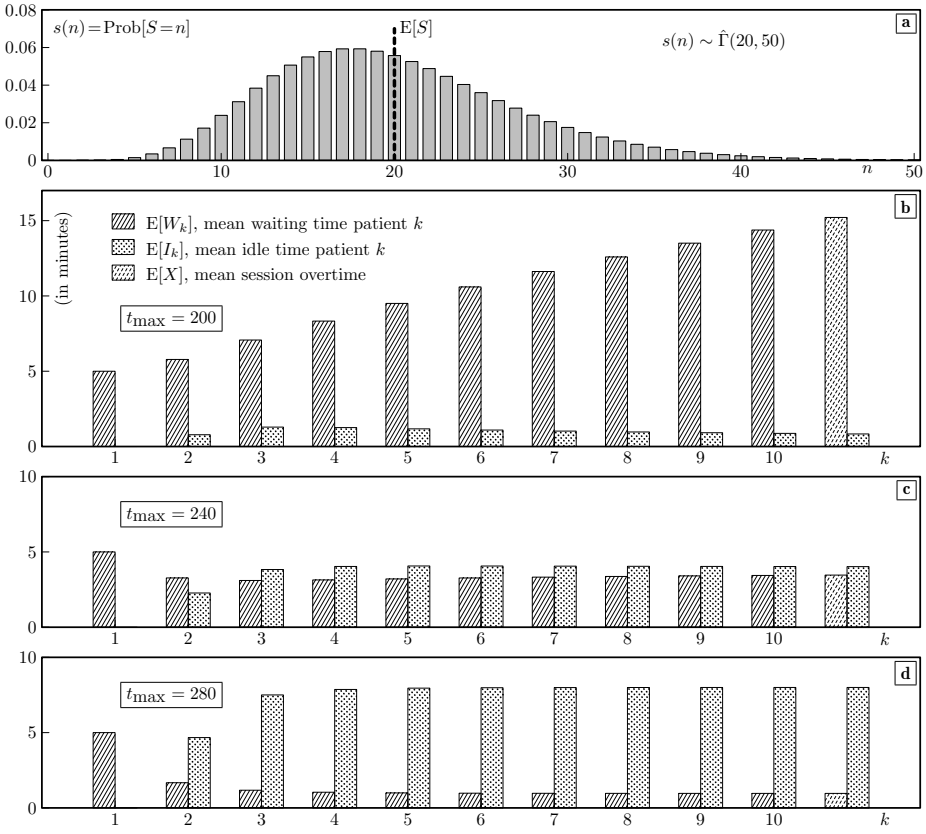

Fig. 1. Evaluation of a schedule with $K = 10$ patients equidistantly spaced in a session of length $t_{\max} = 200$, 240 and 280 minutes. All patients have the same $\hat{\Gamma}(20, 50)$ consultation time distribution with the pmf shown in (a). The physician starts $\theta = 5$ minutes after the session starts.

minute. If $\Gamma(\mu, \sigma^2)$ denotes the (continuous) Γ-distribution with mean μ and variance σ^2, then we refer to the corresponding discrete distribution obtained by (1) as $\hat{\Gamma}(\mu, \sigma^2)$.

First, we consider a session with $K = 10$ patients all having the same consultation time distribution $\hat{\Gamma}(20, 50)$ of which the pmf is shown in Fig. 1(a). The patients are given appointment times equidistantly spaced in the session, i.e. $\tau_k = \lfloor (k-1)a \rfloor$ with $a = t_{\max}/K$, while the physician arrives 5 minutes late in the session. The mean performance of this schedule is shown in Fig. 1(b), (c) and (d) in case the spacing a is 20, 24 and 28 minutes respectively. Note that in (b) the time *given* to each patient is exactly the expected time *needed* by the patient, i.e. $a_k = \mu_k$. Although one would expect this to be an acceptable strategy, it is clearly not since the waiting times of subsequently scheduled patients increase indefinitely (assuming an infinite session $t_{\max} \to \infty$, $K \to \infty$ and $a_k = \mu$ kept

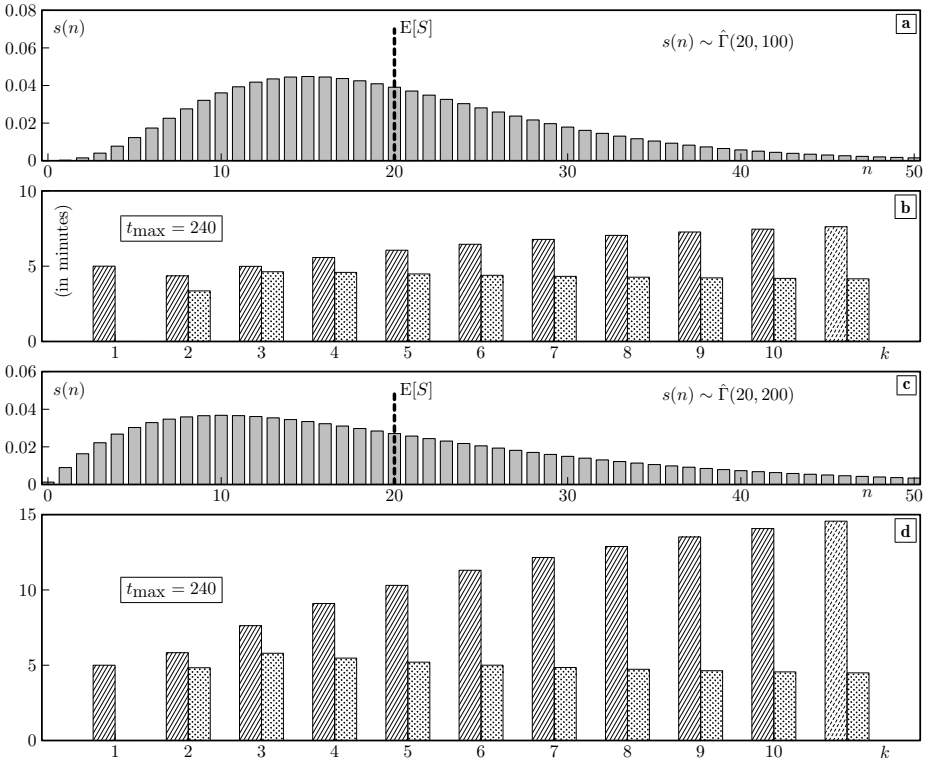

Fig. 2. Schedule with $t_{\max} = 240$, $\theta = 5$ and $K = 10$ equidistant patients. Compared to Fig. 1(c), the variance of the consultation times is increased to 100 in (a)–(b) and to 200 in (c)–(d).

constant), as already observed in [4]. For long sessions it is therefore necessary to choose a larger spacing, for example as $a_k = \mu_k + h\sigma_k$ with some parameter $h > 0$ [8]. Taking the spacing a too large however, as in (d), results in very high physician idle times.

In Fig. 2 we illustrate the influence of consultation time variability in another way. We consider again the schedule of Fig. 1(c) where the patients are placed every 24 minutes but now increase the variance σ of the consultation times from $\sigma = 50$ to $\sigma = 100$ and 200 in Fig. 2(a)–(b) and (c)–(d) respectively, keeping the mean consultation time at 20 minutes. Observe that a high-variance consultation time attributes more uncertainty to the schedule and deteriorates *both* the mean waiting times and idle times [3,5,13,20]. Moreover, the mean waiting times of subsequent patients in Fig. 2(d) increase towards the end of the session, similar as in Fig. 1(b). Here however, if the session were infinite, the waiting times would converge to a limiting value since $a_k < \mu_k$.

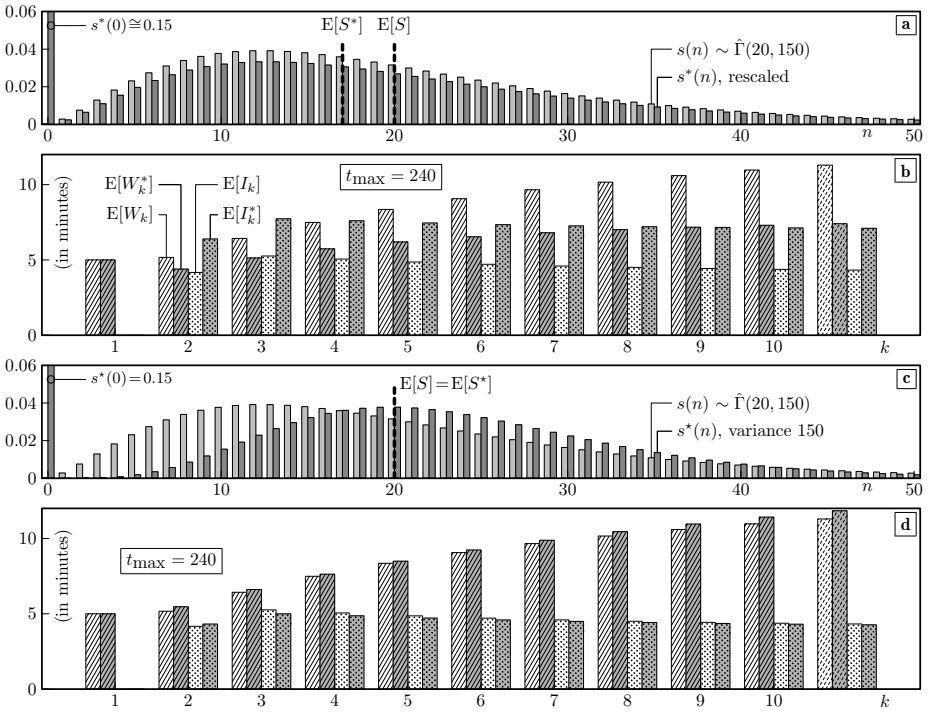

Fig. 3. Schedule with $t_{\max} = 240$, $\theta = 5$ and $K = 10$ equidistant patients. Starting from a $\hat{\Gamma}(20, 150)$ consultation time pmf $s(n)$, we show the effect of adjusting for a no-show probability of 15%. In (a)–(b), the probabilities $s(n)$ are simply rescaled, while in (c)–(d) the shape and scale parameter of the Gamma distribution are adjusted to yield the same mean and variance as for $s(n)$.

In Fig. 3 we illustrate the consequence of no-shows on the schedule's performance, again assuming $t_{\max} = 240$, $\theta = 5$ and $K = 10$ equally spaced patients, i.e. every $a_k = 24$ minutes. The consultation times are all $\hat{\Gamma}(20, 150)$ distributed. In (a) we show the effective pmf $s^*(n)$ obtained from rescaling $s(n)$ as in (2) in order to account for a no-show probability p of 15%. Note that the probability mass $s^*(n) \cong 0.15$ of a zero-length consultation is not shown completely. In (b) the schedule's performance is shown both in the original case where all patients show up and in case the rescaled pmf $s^*(n)$ is used. As the mean consultation time drops from $\mathrm{E}[S] = \mu_k = 20$ to $\mathrm{E}[S^*] = \mu_k^* = 17$ minutes due to the no-shows, the waiting times are lower while the idle times are higher.

In Fig. 3(c)–(d) we illustrate that the consultation time distribution may have an influence beyond its first two moments, due to the sums (9) and (11). Here, both S and S^* have their mean and variance equal to 20 and 150 respectively, although only S is $\hat{\Gamma}$-distributed. The second distribution is obtained by imposing a no-show probability $s^*(0) = p = 15\%$ and choosing a $\hat{\Gamma}$-shape for the other mass points $s^*(n)$, $n > 0$.

3 Assisted Sequential Scheduling

3.1 Cost Function

The 'quality' of a schedule depends on the importance attributed to certain aspects of its performance. For example, it is a usual goal to keep the expected patient waiting times as low as possible, but not at the expense of an excessive physician idle time or session overtime. In most situations, to have the physician standing idle for 1 minute is deemed far less desirable than to keep a patient waiting for the same time period. In general, such differences in appreciation can be represented as a function of the performance criteria we have obtained in Section 2. For example, a suitable 'cost' function C associated with a particular schedule might be

$$C = f(\overline{W}, \overline{I}, \mathrm{E}[X], \ldots)\,, \tag{17}$$

where $\overline{W}$ and $\overline{I}$ respectively are

$$\overline{W} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}[W_k]\,, \qquad \overline{I} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}[I_k]\,.$$

Thus formalised, the 'best' or optimal schedule is that which minimises the cost C. Other criteria such as the number of patients K, the physician lateness θ or variances of waiting and idle times can also be taken into account, as well as specific costs related to particular (types of) patients. Clearly, the results obtained by optimalisation studies highly depends on what criteria are effectively considered in (17). For example, C depends only on $\overline{W}$ and $\overline{I}$ in [3,17,20,13], on $\overline{W}$ and the mean effective session duration $\tau_K + \mathrm{E}[W_K]$ in [22,12], and on $\overline{W}$, $\overline{I}$, $\mathrm{E}[X]$ in [6,14].

3.2 Sequential vs. Advance Scheduling

As mentioned earlier, in appointment scheduling for outpatients, two methods can be distinguished. The first method is to fix the appointment immediately when the patient first calls in. This is called *sequential* scheduling [18,7], because the appointments are fixed one after the other, as the patients call in. Let $\tau(k)$ and $S(k)$ be the appointment time and consultation time of the kth calling patient respectively. If the schedule is already fixed for k patients, the decision at which time $\tau(k+1)$ to schedule the next calling patient must then be based on the schedule so far, as well as on the distribution of $S(k+1)$. Possibly, some global prospect of the future calling patients can be taken into account as well, in case such information is available or can be estimated. The most straightforward manner of sequential scheduling is to put the first calling patient at the start of the session, the next a short time later, and so on, until the end of the session is reached. This results in

$$0 \leqslant \tau(1) \leqslant \tau(2) \leqslant \ldots \leqslant \tau(K) \leqslant t_{\max}\,, \tag{18}$$

and therefore $\tau(k) = \tau_k$. This is called First-come-first-appointment (FCFA) in [20]. However, one can also decide not to maintain the calling order in the schedule, i.e. $\tau(k) \neq \tau_k$, and put a new patient before a previously scheduled patient. This can be due e.g. to preference of the involved patient or because of other imposed time constraints. In any case, the problem with scheduling sequentially is that once a patient is assigned an appointment time it cannot be altered afterwards. If in hindsight it turns out that some minor adjustments would result in lower expected cost C, it is impossible to implement this.

This problem can be overcome by using a two-step procedure. First the desk administrator takes in the calls from all patients, without yet giving them an exact appointment time $\tau(k)$. After enough (say K) patients have called for an appointment, the optimal schedule is determined, either by hand or by means of an automated search algorithm. Once the schedule is decided, each patient is contacted again and notified its appointment time τ_k in the session. This method is referred to as *advance* scheduling, and will usually lead to a better schedule and lower cost C because it is optimised over all decision variables simultaneously, using the best available information on the consultation times of the involved patients. Formally, given K and $t_{\max}$ the optimalisation amounts to a nonlinear integer programming problem with decision variables θ and $\tau(k)$, $k=1,\ldots,K$ in the range $[0, t_{\max}]$. The objective is to determine the integer values that minimise the function (17) and are possibly subjected to some further constraints, e.g. on the range of τ_k imposed by the patient or on the range of θ by the physician. Note that although the decision space is finite, it may be extremely large. For example, assuming that $\theta = 0$ and $\Delta = 1$ minute, an exhaustive search to the optimal schedule for 20 patients on a 4 hour session still requires $241^{20} = 4.3 \cdot 10^{47}$ schedule evaluations. Even in case the order is irrelevant and (18) can be maintained, for example if all patients have the same consultation time distribution, the number of schedules is still $\binom{260}{20} = 3.8 \cdot 10^{29}$. Another disadvantage of advance scheduling is clearly the additional administrative effort of having to contact patients twice.

3.3 Visualisation of Mean Performance

For now we restrict ourselves to the case of sequential scheduling and see how this process can be assisted as much as possible. Suppose a schedule for a particular session already exists when a new patient calls in. This patient must be added to the schedule by the administration in one of the $t_{\max}+1$ slots. In doing so it would be useful to know, for each slot, the waiting time that the patient will experience if it were to be scheduled in that slot, as well as the incurred idle time for the physician. This information can be obtained from the schedule so far as follows, assuming $\theta = 0$ for clarity of reasoning.

Consider again the schedule as defined in the previous section, but disregard patients k+1 to K. That is, consider only the consultations of the first k patients with appointment times τ_1 to τ_k. Let $R_{k,i}$, $0 \leqslant i \leqslant t_{\max} - \tau_k$, denote the *remaining work* for the physician i slots after τ_k, the last patient's appointment. Likewise, define $J_{k,i}$ as the time that the physician has been idle in that slot. We refer to this quantity as the *running idle time*. With the auxiliary variable

$$Q_{k,i} = W_k + S_k - i\,, \tag{19}$$

we have

$$R_{k,i} = (Q_{k,i})^+\,, \quad \text{and} \quad J_{k,i} = (-Q_{k,i})^+\,, \tag{20}$$

analogous to (4)–(5). Clearly, $R_{k,i}$ and $J_{k,i}$ respectively correspond to the waiting time and idle time of an additional patient if it were to be scheduled in slot τ_k+i, which is precisely the reason why these quantities are useful in sequential scheduling. Their moments can be calculated in exactly the same way as in (8)–(11) and (14)–(16). We thus find

$$\begin{aligned} \mathrm{E}[R_{k,i}] &= \mathrm{E}[W_k] + \mu_k - i + \bar{\ell}_{k,i}\,, \\ \mathrm{Var}[R_{k,i}] &= \mathrm{Var}[W_k] + \sigma_k^2 + \bar{\ell}_{k,i}^2 - \bar{\bar{\ell}}_{k,i} - 2\bar{\ell}_{k,i}\mathrm{E}[R_{k,i}]\,, \end{aligned} \tag{21}$$

and

$$\mathrm{E}[I_{i,k}] = \bar{\ell}_{k,i}\,, \qquad \mathrm{Var}[I_{k,i}] = \bar{\bar{\ell}}_{k,i} - \bar{\ell}_{k,i}^2\,, \tag{22}$$

with

$$\bar{\ell}_{k,i} = \sum_{m=0}^{i} \sum_{n=0}^{i-m} (i-n-m) s_k(m) w_k(n)\,, \tag{23}$$

$$\bar{\bar{\ell}}_{k,i} = \sum_{m=0}^{i} \sum_{n=0}^{i-m} (i-n-m)^2 s_k(m) w_k(n)\,. \tag{24}$$

For example, the expected remaining work and running idle time for a session of length $t_{\max} = 120$ with 6 already scheduled patients is visualised in Fig. 4. This graph shows the session on the horizontal axis and where the 6 patients are scheduled. For each patient k, $\mathrm{E}[W_k]$ and $\mathrm{E}[I_k]$ are plotted as a bar on the positive and negative vertical axis respectively, which shows the same information as in Figs. 1–3. Additionally, the black bar shows the mean consultation time which is the average workload each patient poses on the system. The graph also gives an idea of how the schedule performs in *between* appointments by showing the curves of the mean remaining work and running idle time. That is, for each slot $t \geqslant \tau_k$ we show the average amount of time $\mathrm{E}[R_{k,t-\tau_k}]$ that the physician still has to spend in slot t on the first k patients. Likewise, the curves on the negative vertical axis show the average amount of time $\mathrm{E}[J_{k,t-\tau_k}]$ since the physician completed the consultation of patient k. Clearly, as the session progresses, the remaining work decreases while the running idle time increases. Note also that the remaining work due to the first k patients at the appointment time of the following patient coincides with that patient's waiting time and likewise for idle times, i.e.

$$R_{k,a_k} = W_{k+1}\,, \quad \text{and} \quad J_{k,a_k} = I_{k+1}\,, \tag{25}$$

at least if $\tau_k \geqslant \theta$. This follows directly from the fact that (19)–(20) coincides with (4)–(5) for $i = a_k$. In the same way, for $i = 0$ in (19)–(20) we have

$$R_{k,0} = W_k + S_k\,, \quad \text{and} \quad J_{k,0} = 0\,, \tag{26}$$

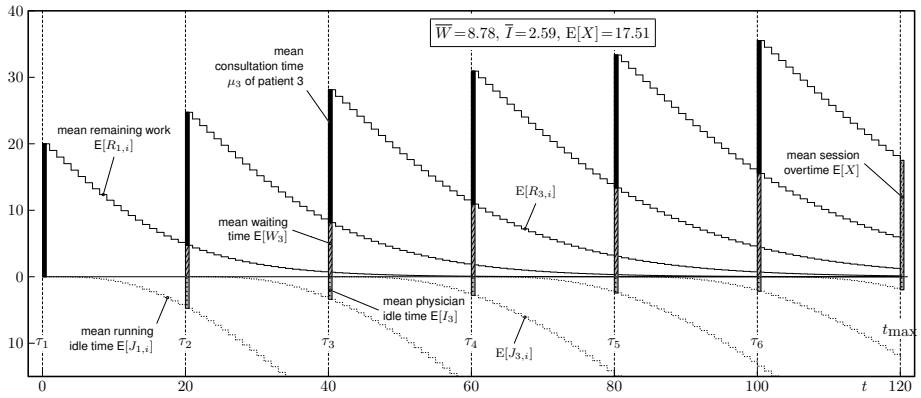

Fig. 4. Visualisation of a schedule with $t_{\max}=120$ slots, $\theta=0$ and $K=6$ equidistant patients. For each patient k the mean remaining work $\mathrm{E}[R_{k,i}]$ is plotted and, in the negative vertical axis, the mean running idle time $\mathrm{E}[I_{k,i}]$. The mean consultation, waiting and idle time of each patient is also shown, as well as the expected overtime $\mathrm{E}[X]$ at the end of the session. All patients have the same $\hat{\Gamma}(20,150)$ consultation time distribution. For this session, we have $\overline{W}=8.78$, $\overline{I}=2.59$ and $\mathrm{E}[X]=17.51$.

which, after taking expected values, is also evident from the graph. Although shown in Fig. 4 for demonstration purposes, it is not necessary to calculate $\mathrm{E}[R_{k,i}]$ of the kth patient for $i \geqslant a_k$, that is, beyond the appointment time of the next patient. Let us define the envelope of the mean remaining work curves and running idle time curves respectively as the functions

$$R(t) = \mathrm{E}[R_{k^*(t),t-\tau^*(t)}]\,, \quad J(t) = \mathrm{E}[J_{k^*(t),t-\tau^*(t)}]\,, \tag{27}$$

with

$$k^*(t) = \max\{k : \tau_k \leqslant t\}\,, \quad \tau^*(t) = \tau_{k^*(t)}\,. \tag{28}$$

In the examples of Figs. 5 and 6 discussed below, the curves $R(t)$ and $J(t)$ on positive and negative vertical axis will be shown with a solid and dotted line respectively. Such a visual representation of the schedule's average performance can assist the administrator if a new patient needs an appointment. For each slot t, the administrator can immediately see what the mean waiting and idle time for the new patient will be if it is appointed to that slot. If the new patient is effectively appointed to some slot t, everything from slot t onwards must be recalculated in order to visualise the new situation.

3.4 Examples

Let us consider a scenario where the patients are heterogeneous and have consultation times according to either of the four following distributions: (a) uniform between 5 and 15 minutes, (b) Poisson with a mean of 15 minutes, (c) $\hat{\Gamma}(20,200)$

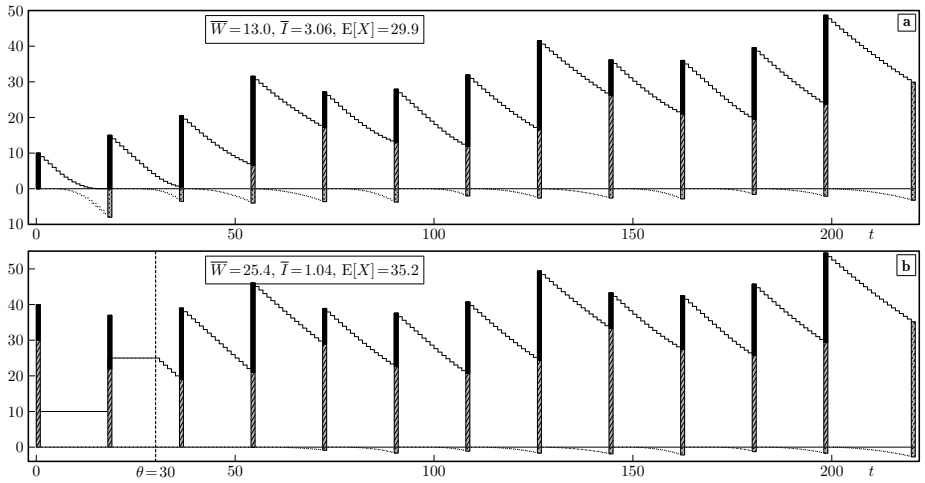

Fig. 5. Visualisation of a 4-hour session with 12 patients: remaining work $R(t)$ and running idle time $J(t)$ envelopes. The consultation times of the patients alternate between the four mentioned distributions. In (a), the patients are spaced at 18 minutes, which is about the overall mean of their consultation times. In (b) the physician arrives half an hour late.

and (d) geometric with a mean of 25 minutes. These distributions have expected values 10, 15, 20, 25 and variances 10, 15, 200, 650 respectively. Let us assume that subsequent scheduled patients have consultation times circulating between these four distributions, i.e. S_1 is distributed according to (a), S_2 to (b), S_3 to (c), S_4 to (d), S_5 to (a) again and so on. If it is not known to which type the consultation of a patient belongs to, we may assume that each type is equally likely which results in a pmf $s(n) = \frac{1}{4}\sum_{k=1}^{4} s_k(n)$, $n \geqslant 0$, with mean 17.5 and variance 250.

In Fig. 5 the visualisation of a 240-minute session is shown for $K = 12$ patients. The administrator here does not take into account the heterogeneity of the patients and only uses the overall consultation time distribution as information. Patients are given appointments 18 minutes apart, which equals $\lceil \mathrm{E}[S] \rceil$. So, even if for example the first patient cannot take more than 15 minutes, it is given an interval of 18 minutes. In (b) the same schedule is shown, but now the physician arrives 30 minutes after the start of the session. Although $\overline{W}$ increases considerably this way, the mean idle time $\overline{I}$ is effectively reduced from 3.06 to 1.04 minutes. Note also that even though the physician starts 30 minutes later, the expected session overtime increased by only 5.3 minutes.

In Fig. 6 on the other hand, the administrator uses knowledge about the patient's consultation type for scheduling. In (a) Bailey's rule is applied: two patients are scheduled in the first slot while the rest are given an interval equal to their mean consultation time, $a_1 = 0$, $a_k = \lceil \mathrm{E}[S_k] \rceil$, $k = 2, \ldots, K-1$. In (b), we show an appointment rule that is particularly easy to apply if the envelopes

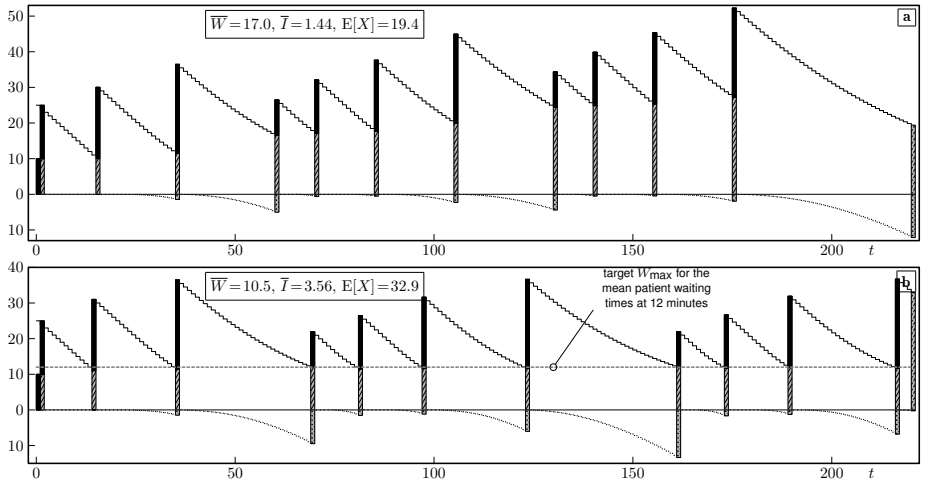

Fig. 6. Contrary to Fig. 5 the administrator knows which one of the 4 distributions each patient has. In (a) Bailey's rule is followed, while in (b) all patients are targeted to have an expected waiting time of less than 12 minutes.

$R(t)$ and/or $J(t)$ are available. Suppose that our main optimalisation goal is *fairness* among patients, i.e. we want the mean waiting times $\mathrm{E}[W_k]$ to be more or less equal such that no patient is favoured by its position in the session. To achieve this, we can impose a target value $W_{\max}$ for the mean patient waiting times of, say, 12 minutes which is indicated by the horizontal grey line. The point where this line intersects with the remaining work envelope $R(t)$ is where the next patient will be scheduled. There is no sense in making the first patient wait however, which makes that $\overline{W} = 10.5$ minutes instead of 12 minutes. This way of sequential scheduling could be extended by an additional target $I_{\max}$ for the idle times. If k patients are already scheduled, the appointment time of patient $k+1$ is decided by

$$\tau_{k+1} = \min\{t_W, t_I\}\,,$$

with

$$t_W = \min\{t \geqslant \tau_k \,:\, R(t) < W_{\max}\}\,, \quad \text{and} \quad t_I = \max\{t \,:\, J(t) < I_{\max}\}\,.$$

4 Conclusions

We have proposed an analytic approach for evaluating appointment schedules on finite sessions that allows to obtain accurate results with very low computational complexity. By imposing a discrete-time setting and using Lindley's recursion, we show that only a limited set $\mathcal{W}$ of waiting time probabilities need to be calculated in order to obtain the moments of waiting and idle times of the patients appointed to the session. The fact that the evaluation can be done very

fast, makes our approach an ideal candidate for use in optimalisation studies. Additionally, we propose two new metrics that can assist in scheduling the patients sequentially: the mean remaining work and mean running idle time envelopes, which characterise the average performance of the schedule in each slot of the session.

Acknowledgements

The third author is a postdoctoral fellow with the Fund for Scientific Research – Flanders (FWO-Vlaanderen).

References

1. Alexopoulos, C., Goldsman, D., Fontanesi, J., Kopald, D., Wilson, J.R.: Modeling patient arrivals in community clinics. Omega 36, 33–43 (2008)
2. Babes, M., Sarma, G.: Out-patient queues at the Ibn-Rochd health centre. The Journal of the Operational Research Society 42(10), 845–855 (1991)
3. Bailey, N.T.J.: A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. Journal of the Royal Statistical Society. Series B (Methodological) 14(2), 185–199 (1952)
4. Bailey, N.T.J.: A note on equalising the mean waiting times of successive customers in a finite queue. Journal of the Royal Statistical Society. Series B (Methodological) 17(2), 262–263 (1955)
5. Cayirli, T., Veral, E.: Outpatient scheduling in health care: A review of literature. Production and Operations Management 12(4), 519–549 (2003)
6. Cayirli, T., Veral, E., Rosen, H.: Assessment of patient classification in appointment system design. Production and Operations Management 17(3), 338–353 (2008); Conference of the Production-and-Operations-Management-Society and the College-of-Service-Operations, New York (October 2004)
7. Chakraborty, S., Muthuraman, K., Lawley, M.: Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Transactions 42(5), 354–366 (2010)
8. Charnetski, J.: Scheduling operating room surgical procedures with early and late completion penalty costs. Journal of Operations Management 5(1), 91–102 (1984)
9. Denton, B., Gupta, D.: A sequential bounding approach for optimal appointment scheduling. IIE Transactions 35, 1003–1016 (2003)
10. Green, L.V., Savin, S.: Reducing delays for medical appointments: A queueing approach. Operations Research 56(6), 1526–1538 (2008)
11. Harper, P., Gamlin, H.: Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. OR Spectrum 25(2), 207–222 (2003)
12. Hassin, R., Mendel, S.: Scheduling arrivals to queues: A single-server model with no-shows. Management Science 54(3), 565–572 (2008)
13. Ho, C.J., Lau, H.S.: Minimizing total-cost in scheduling outpatient appointments. Management Science 38(12), 1750–1764 (1992)
14. Kaandorp, G.C., Koole, G.: Optimal outpatient appointment scheduling. Health Care Management Science 10, 217–229 (2007)

15. Lehmann, T.N.O., Aebi, A., Lehmann, D., Olivet, M.B., Stalder, H.: Missed appointments at a Swiss university outpatient clinic. Public Health 121(10), 790–799 (2007)
16. Lindley, D.: The theory of queues with a single server. Mathematical Proceedings of the Cambridge Philosophical Society 48(2), 277–289 (1952)
17. Liu, L., Liu, X.: Block appointment systems for outpatient clinics with multiple doctors. Journal of the Operational Research Society 49(12), 1254–1259 (1998)
18. Muthuraman, K., Lawley, M.: A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Transactions 40(9), 820–837 (2008)
19. Reinus, W.R., Enyan, A., Flanagan, P., Pim, B., Sallee, D.S., Segrist, J.: A proposed scheduling model to improve use of computed tomography facilities. Journal of Medical Systems 24(2), 61–76 (2000)
20. Rohleder, T.R., Klassen, K.J.: Using client-variance information to improve dynamic appointment scheduling performance. Omega, International Journal of Management Science 28(3), 293–302 (2000)
21. Sola-Vera, J., Saez, J., Laveda, R., Girona, E., Garcia-Sepulcre, M.F., Cuesta, A., Vazquez, N., Uceda, F., Perez, E., Sillero, C.: Factors associated with non-attendance at outpatient endoscopy. Scandinavian Journal of Gastroenterology 43(2), 202–206 (2008)
22. Wang, P.P.: Optimally scheduling n customer arrival times for a single-server system. Computers & Operations Research 24(8), 703–716 (1997)
23. Zonderland, M.E., Boer, F., Boucherie, R.J., de Roode, A., van Kleef, J.W.: Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. Anesthesia and Analgesia 109(5), 1612–1621 (2009)